

Enterprise Taxonomies – Type, Integration & Design Issues

Denise A. D. Bedford, Ph.D.

Senior Information Officer
Information Solutions Group
The World Bank Group

December 8, 2004

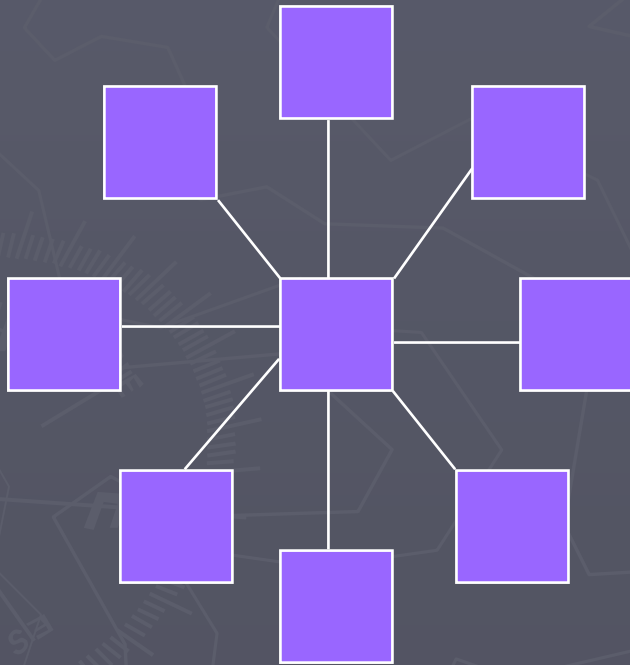
Information Management System Architectures

- ▶ The underlying architecture of an information management system is complex
- ▶ Taxonomies are essential structures in all information management systems
- ▶ In order to build an efficient, sustainable information architecture, we should
 - Understand the different kinds of taxonomies
 - Have sufficient familiarity with their purpose to select the right kind of taxonomy for an application

Taxonomy Basics

- ▶ There are four types of taxonomies
 - Faceted
 - Flat
 - Hierarchical
 - Network
- ▶ Some are explicit/visible, others are implicit/invisible
- ▶ There are significant design consideration when implementing each different type
- ▶ Let's review each quickly

Facet Taxonomies



Faceted taxonomy represented as a star data structure. Each node in the star structure is linked to the center focus. Any node can be linked to other nodes in other stars. Appears simple, but becomes complex quickly.

Type 1: Faceted Taxonomy

- ▶ There are no inherent relationships among categories in a faceted taxonomy – like a flat taxonomy
- ▶ All categories in a faceted taxonomy relate to a single object -
- may describe a property or a value, different views or aspects of a single topic
- ▶ The primary implicit application of faceted taxonomies today & historically is as implicit metadata records
- ▶ Today, portals and e-business systems are primary metadata users
- ▶ Bank standard metadata

Designing Faceted Taxonomies

- ▶ Characteristics of each facet should be defined fully and distinctly -- while all facets pertain to a common object, each has a distinct behavior
- ▶ Users should be able to manipulate facets distinctly -- it is important to define each facet exclusively, without overlap with other facets
- ▶ Each facet may be managed or governed by another kind of taxonomy

What is metadata?

- ▶ Metadata (or "data about data") describe the content, quality, condition, and other characteristics of data.
- ▶ Metadata are used to organize and maintain investments in data, to provide information to search engines, data catalogs and clearinghouses, to manage information assets, and to aid data transfers

What is metadata?

- ▶ Metadata (or "data about data") describe the content, quality, condition, and other characteristics of data.
- ▶ Metadata are used to organize and maintain investments in data, to provide information to search engines, data catalogs and clearinghouses, to manage information assets, and to aid data transfers

Purpose of Bank Metadata

Identification/
Distinction

Search &
Browse

Use Management

Compliant Document
Management

Agent	Country	Authorized By	Record Identifier
Title	Region	Rights Management	Disposal Status
Date	Abstract/ Summary	Access Rights	Disposal Review Date
Format	Keywords	Location	Management History
Publisher	Subject-Sector- Theme-Topic	Use History	Retention Schedule/Mandate
Language	Business Function		Preservation History
Version			Aggregation Level
Series & Series #			Relation
Content Type			

Metadata Capture Methods

Identification/
Distinction

Search &
Browse

Use Management

Compliant Document
Management

Agent	Country	Authorized By	Record Identifier
Title	Region	Rights Management	Disposal Status
Date	Abstract/ Summary	Access Rights	Disposal Review Date
Format	Keywords	Location	Management History
Publisher	Subject-Sector- Theme-Topic	Use History	Retention Schedule/Mandate
Language	Business Function		Preservation History
Version			Aggregation Level
Series & Series #			Relation
Content Type			

Human Capture

Programmatic Capture

Extrapolate from Business
Rules

Inherit from System Context

Facet Taxonomy as Foundation

- ▶ The best approach for integrating into a faceted taxonomy is to work across the sources that will be contributing content to the taxonomy
- ▶ Look around the enterprise – do you have different faceted taxonomies to integrate?
- ▶ Begin by analyzing the facets in each structure -- this becomes the 'encoding scheme' or your basic data structure
- ▶ You will need to first harmonize at the individual facet level across sources

Facet Taxonomy as Foundation

- ▶ Analyzing facets is a labor intense task that often requires consulting system data dictionaries, data models or references sources in order to understand...
 - definition and purpose
 - specifications – element size, fixed or variable, dependencies, subelements, etc.
 - behavior in the system – controlled or free values, controlling source, etc.
 - basic structure – each facet may have a different kind of structure – flat, hierarchical, faceted, network

Faceted Taxonomy as Foundation

- ▶ In your facet taxonomy, make sure your facets are orthogonal (do not overlap in intent and coverage) at the top
- ▶ Then begin to work on the definition, structure, behavior and values for each facet in the taxonomy
- ▶ You start the analysis anew for each facet – ask all the basic questions again, facet by facet
- ▶ A facet in a faceted taxonomy can have a hierarchical taxonomy as a governing structure

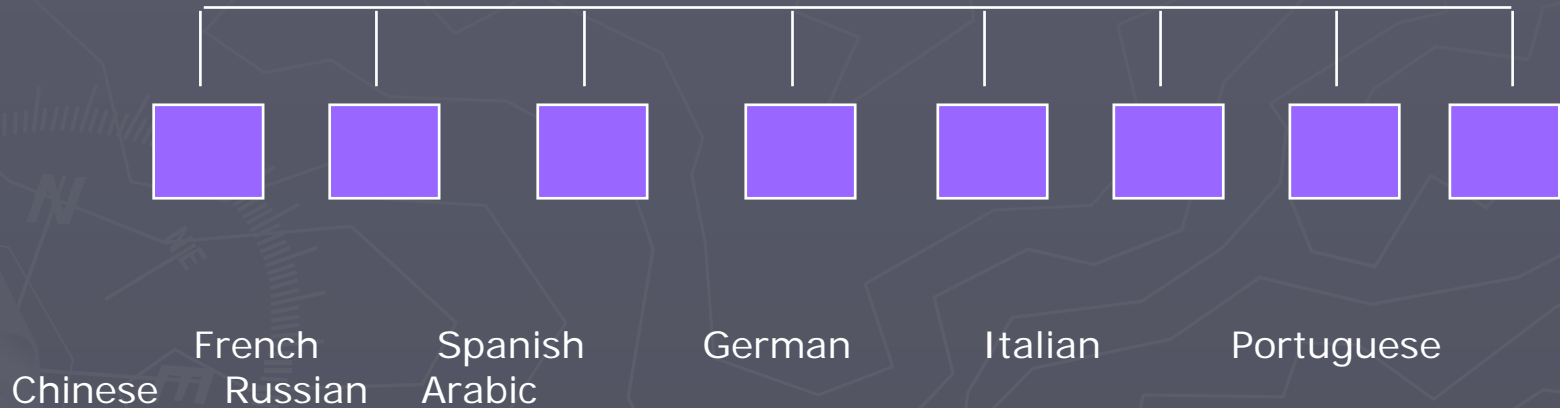
Select or Create a Facet Taxonomy?

- ▶ Use of standard metadata schemes facilitates interoperability by allowing metadata records to be exchanged and imported into other systems that use the same scheme
- ▶ These mappings, or *crosswalks*, help users of one scheme to understand another, can be used in automatic translation of searches, and allow records created according to one scheme to be converted by program to another
- ▶ If a locally created metadata scheme is used in preference to a standard scheme, a crosswalk to some standard scheme should be developed

Selecting a Facet Taxonomy

- ▶ Published metadata schemes available to choose from -- Dublin Core, GILS, COSATI, UDDI, SCORM
- ▶ Select a scheme that is appropriate to:
 - your information & data
 - level of resources you can devote to metadata
 - level of expertise of the metadata creators
 - the expected use and users of the collection
 - the granularity of description, that is, whether to create descriptive records at the collection level, at the item level, or both, in light of the desired depth and scope of access to the materials
 - The level of interoperability you need among collections.

Flat Taxonomy Structure



Type 2: Flat Taxonomies

- ▶ Flat taxonomies group content into a controlled set of categories
 - no inherent relationship among the categories in a flat taxonomy -- they are co-equal members of a single structure
 - can move from one category to another without having to think about the relationship between them
 - concept of a *flat* taxonomy may be counter intuitive to some
- ▶ Consider how often you use flat taxonomies everyday
 - alphabetical listings of people in a directory of expertise
 - a pull-down menu of country names or geographical regions
 - simple alphabetical listings of product groupings

Designing Flat Taxonomies

- ▶ Flat taxonomies are easy to create
- ▶ Flat taxonomies do not require complex interface design and extensive usability testing
- ▶ We have learned from usability engineers how to implement flat taxonomies
 - Flat taxonomies used for explicit information structures generally should consist of 30 or fewer categories;
 - More than 30 categories may be presented in a flat taxonomy, if the categories are intuitive to users (i.e. lists of countries, states, languages, etc.);

Implicit Flat Taxonomies

- ▶ Language codes & names – ISO standard
- ▶ Format types – ISI standard types
- ▶ Rights management values (simple picklist)
- ▶ Information disclosure status values (simple picklist)
- ▶ Security classification scheme values (simple picklist)

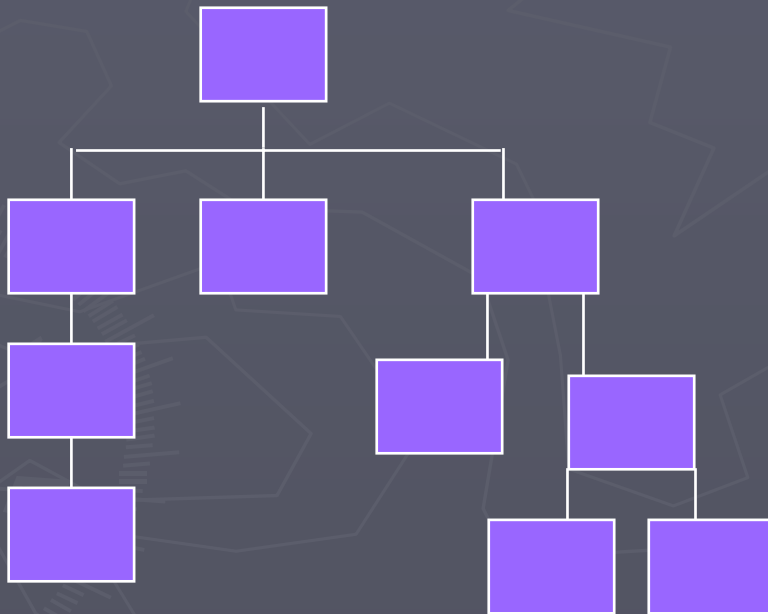
Flat Taxonomy as Foundation

- ▶ You have a simple, one level list
- ▶ All content fit into that one list without confounding
- ▶ The behavior is simple – users simply select & display all content or functionality at one level

Building an integrated flat taxonomy

- ▶ Look across all values to find categories
- ▶ Start by reviewing all of the categories you currently have – don't reinvent when you don't have to
- ▶ Review the 'goodness of fit' of all the content to be integrated to the categories
- ▶ Try to name the clusters and see if they make sense to users
- ▶ If you cannot fit all of your content into the simple one-level list, you should not use a flat taxonomy

Hierarchical Taxonomy



A hierarchical taxonomy is represented as a tree data structure in a database application. The tree data structure consists of nodes and links. In an RDBMS environment, the relationships become associations. In a hierarchical taxonomy, a node can have only one parent.

Type 3: Hierarchical Taxonomies

- ▶ Group content into two or more levels
- ▶ Resemble tree structures when they are fully elaborated
- ▶ Hierarchical categories typically have only one broader or parent category
- ▶ Bank Major Sectors/Sectors, Major Themes/Themes

Type 3: Hierarchical Taxonomies

- ▶ Relationships among categories in hierarchical taxonomies have particular meaning
 - Relationship between a top level category & subcategory may mean group membership or refinement of the top category by a particular characteristic or feature
 - Moving up the hierarchy means expanding or broadening the category
 - Moving down the hierarchy means refining or qualifying the category

Categorization

- ▶ Categorization technologies classify documents into groups or collections of resources
- ▶ An object is assigned to a category or schema class because it is 'like' the other resources in some way
- ▶ Categorization improves precision of search by letting users search for concepts within a category of resources
- ▶ Different from describing all of the concepts that are substantively treated in an object

Botswana Shashe Infrastructure Project- Example

- ▶ Topic or Sector = Infrastructure or Water Supply & Sanitation (categorization)
- ▶ Keywords = Dams, pumping stations, water pipelines, water treatment plants, coal mine water supply, water distribution systems (concepts)
- ▶ Assigning the project document to the Water Supply & Sanitation sector helps to narrow the category in which to search
- ▶ Assigning concepts to the document makes it possible for a user to search by concepts or keywords for specific knowledge or information contained in the document itself

Bank Hierarchical Taxonomies

- ▶ Primary & Secondary Content Types
- ▶ Enterprise Classification Scheme
- ▶ Business Activity Taxonomy

Designing Hierarchical Taxonomies

► Hierarchical taxonomies should:

- have content at every level -- empty categories present empty value to users
- be less than four levels deep in most cases
- be at least two categories for each branch in the taxonomy -- do not branch for a single category
- be sufficient content in each category to warrant existence
- balance breadth & depth -- users must work harder to use a taxonomy three categories broad & nine deep than to use one that is seven wide and two deep

Designing Hierarchical Taxonomies

- ▶ There is more than one way to implement a hierarchy:
 - Progressive disclosure of layers across sites or pages -- Ebay model
 - Cascading or expanding menus -- United Nations web site
 - Pop-up menus linked to stationary menus -- United Nations web site
 - Category and subcategory labels in a multi-column display -- Nordstrom's second level pages

Hierarchical Taxonomy Design Issues

- ▶ Hierarchical taxonomies should:
 - be balanced across each level of the taxonomy to provide users with a predictable experience
 - be offset with search functions
 - should never be displayed into flat structures
 - be reviewed periodically

Hierarchical Taxonomy as Foundation

- ▶ If you are trying to build collections of information that can be progressively refined without changing the focus
- ▶ If you are trying to control for variations in naming conventions or aliasing

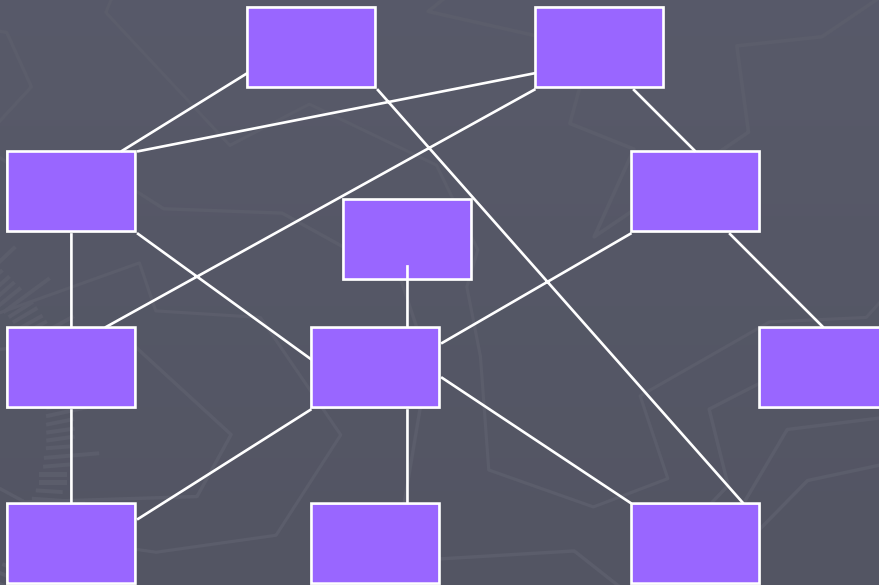
Building an integrated hierarchical taxonomy

- ▶ Start at the top and compare top level categories across hierarchies
 - How many categories are there?
 - Do they have the same level of granularity?
 - Do they have the same 'warrant' or do they represent different views or facets of the same object?
 - If you begin to find facets, stop to consider whether they are always strictly hierarchical – if not, switch back to a facet taxonomy

Building an integrated hierarchical taxonomy

- ▶ Draft a common top level of categories by clustering like categories
 - Review for balance across all top level categories
 - Combine where granularity is too fine, split where granularity is too coarse
- ▶ Define category labels that describe all of the content in the top level and its subcategories
- ▶ Review the amount of content associated with each category
- ▶ Map existing taxonomy categories to your new taxonomy

Network Taxonomies



A network taxonomy is a plex data structure. Each node can have more than one parent. Any item in a plex structure can be linked to any other item. In plex structures, links can be meaningful & different.

Type 4: Network Taxonomies

- ▶ Organizes content into both hierarchical and associative categories
- ▶ May look like a computer network topology
- ▶ Many relationships among categories or “nodes”
- ▶ Relationships may have many different meanings
- ▶ Category may have more than one higher level category
- ▶ Any category in the taxonomy may be linked to any other category

Examples of Network Taxonomies

- ▶ Thesauri, concept maps & semantic networks can be explicit or implicit
- ▶ World Bank Thesaurus – <http://www2.multites.com/wb/>
- ▶ Can be designed transparently into the knowledge management system as:
 - thesaurus facilitated search systems
 - recommender engines (...if you liked this, you might also like this)
 - vocabulary cross-walks from one source system to another
 - topic map cross-walks from one knowledge domain to another
 - Semantic networks

Networked Taxonomy as Foundation

- ▶ Use anytime that the information architecture can be used to move up, down, *and sideways*
- ▶ If you are trying to implement a thesaurus into a search system
- ▶ If you are building a semantic web framework
- ▶ If you are building a vocabulary cross-walk

Building an integrated networked taxonomy

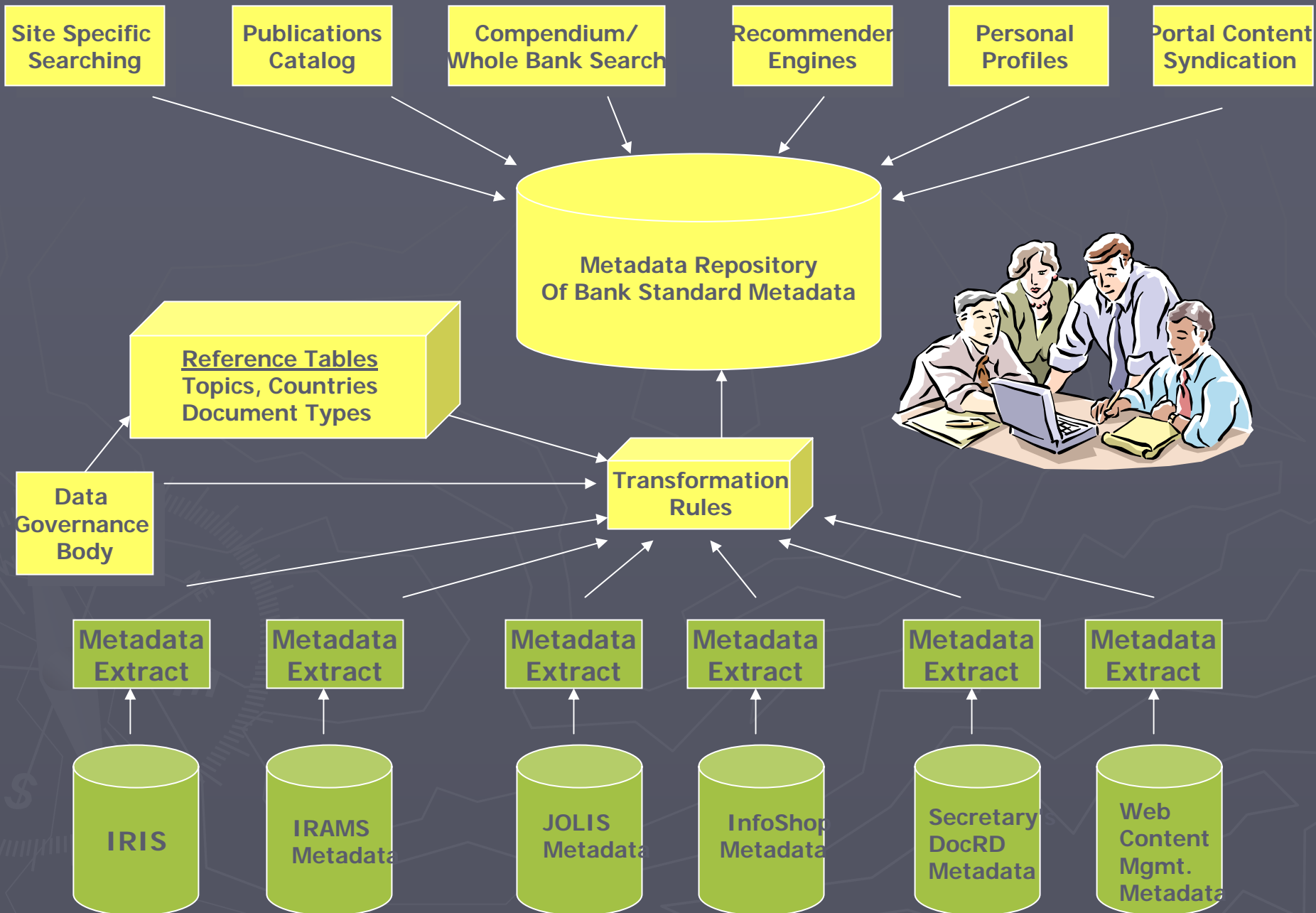
- ▶ Start at the lowest level of the sources you are trying to integrate
- ▶ Create a database or file of all the concepts and note their sources
- ▶ Review all of the concepts you have
 - Discover existing overlaps across sources
 - Discover the links that should be built amongst concepts that come from different sources
 - Determine the kinds of links that should be made

Bank Approach to Metadata



Metadata Warehouse Approach to Metadata & Content Management

- ▶ How to manage complexity, maintain source systems, respect content security & still meet users expectations?
- ▶ Create warehouse of metadata pertinent to access, search & syndication
- ▶ Working from a type of content perspective since metadata converge around kinds of content & within source systems
- ▶ Define attribute super classes to which existing metadata are mapped
- ▶ Attributes may be rationalized, harmonized or value-controlled within super classes



Automated Metadata Capture – Enterprise Metadata Profiles

Enterprise Metadata Strategy

- ▶ Business system metadata suited to business purpose
- ▶ Bankwide & public access requires a different perspective & approach
- ▶ Solution -- map existing business system metadata to metadata superclasses
- ▶ Mapping takes place in a high level metadata warehouse
- ▶ Harmonize at the superclass level
- ▶ Harmonize & integrate using reference sources maintained outside of business systems

Full Taxonomy Integration

The diagram illustrates a hierarchical structure with a central node (dark grey square) connected to several branches. The branches are color-coded and further subdivided:

- Teal Branch (Left):** A teal square connected to the center, which further branches into two teal squares, each of which branches into two more teal squares.
- Purple Branch (Top):** A purple square connected to the center, which further branches into two purple squares, each of which branches into two more purple squares.
- Yellow Branch (Bottom):** A yellow square connected to the center, which further branches into two yellow squares, each of which branches into two more yellow squares.
- Green Branch (Right):** A green square connected to the center, which further branches into two green squares, each of which branches into two more green squares.

The background features a faint, stylized map of the world.

Types of World Bank Enterprise Metadata

- ▶ Identify & distinguish information - Integrity of information
- ▶ Facilitate search & browse - Long term search strategy
- ▶ Manage use of information - Security classification, use rates, copyright compliance, disclosure policy implementation
- ▶ Ensure compliant information management - retention & dispositioning of information, ensuring records are archived & protected, ensuring that convenience copies are discarded

World Bank Taxonomy Structures

- ▶ Each facet is supported by different types of taxonomy structures
 - simple lists of controlled values
 - alias hierarchies of successor/predecessor terms
 - hierarchy topic maps
 - complex network structures - thesauri

Metadata Reference Sources

- ▶ Reference sources:
 - centrally managed
 - enforce quality control over metadata
 - manage change in values over time
 - harmonize institutional collection values without impacting the collection itself
- ▶ Represented as Oracle data classes - long term vision is for institutional collections to retrieve/use reference sources
- ▶ Use ER win software to manage data classes and BPwin to maintain the data flows and business rules